DOCUMENT RESUME

ED 458 279 TM 033 443

AUTHOR Morris, Don R.
TITLE Causal Inference and System Dynamics in Social Science Research: A Commentary with Example.
PUB DATE 2001-07-00
NOTE 28p.; Paper presented at the International Conference of the System Dynamics Society (19th, Atlanta, GA, July 23-27, 2001).
PUB TYPE Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Causal Models; *Educational Research; *Research Methodology; *Social Science Research
IDENTIFIERS *Causal Inferences; *System Dynamics

ABSTRACT

The focus of this paper is causal inference in social and educational research. A concern with causality has had a profound impact on the kinds of questions that may be addressed in research, on how they must be formulated, and on the methodology that must be applied. In the social sciences the prevailing experimental paradigm is used to address causal inference in a way that has forced a comparative or counterfactual approach, to the exclusion of physical cause. Recently, Lawrence Mohr has proposed a means of putting physical cause on an equal footing with the counterfactual, an achievement with important implications. In this paper, Mohr's concepts are joined with the resources of system dynamics. Following a discussion of the concepts, an example from educational research based on physical causal reasoning and system dynamics methodology is presented and the approach assessed. (Contains 5 figures and 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 279

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

D. Morris

CAUSAL INFERENCE AND SYSTEM DYNAMICS IN SOCIAL SCIENCE RESEARCH: A COMMENTARY WITH EXAMPLE

Don R. Morris
Miami-Dade County Public Schools

Presented at the
19th International Conference of the System Dynamics Society

Atlanta, Georgia
July, 2001

TM033443

BEST COPY AVAILABLE

Causal Inference and System Dynamics in Social Science Research: A Commentary with Example

Don R. Morris
Miami-Dade County Public Schools

The focus of this paper is causal inference in social and educational research. A concern with causality has had a profound impact on the kinds of questions that may be addressed in research, on how they must be formulated, and on what methodology must be applied. In the social sciences the prevailing experimental paradigm is used to address causal inference in a way that has forced a comparative or counterfactual approach, to the exclusion of physical cause. Recently, Lawrence Mohr has proposed a means of putting physical cause on an equal footing with the counterfactual, an achievement with important implications. In this paper, Mohr's concepts are joined with the resources of system dynamics. Following a discussion of the concepts, an example from educational research based on physical causal reasoning and system dynamics methodology is presented and the approach assessed.

Keywords: research methodology; causality; educational policy; physical cause; modus operandi

Social scientists and system dynamicists approach research from very different perspectives. The source of the contention has ordinarily been identified as feedback or systems thinking, but as serious and obvious as they are, feedback and systems thinking are not the only factors setting system dynamics apart from the social sciences. There appears also to be a fundamentally different understanding of the way in which cause can be validly inferred.[1]

I contend that causality is an important and underestimated issue preventing a common dialogue from being established. My focus in this paper is on the fact that a concern with causality has had a profound impact on what kinds of questions may be addressed in social research, on how they must be formulated, and on what methodology must be applied. This paper proceeds from two points. First, the experimental paradigm prevails in the social sciences, and it is used to address the "fundamental problem of causality" (causes cannot be observed) in a way that has forced a comparative or correlational approach to causal inference, without alternative. Second, Lawrence Mohr (1996) claims to have found a means of including the concept of physical cause on an equal footing with that which now prevails. It is my intention to show how Mohr's concepts, when joined with the resources of system dynamics, can contribute an attractive alternative in the social sciences, broadening the scope of research. I will first discuss in the abstract causal inference and the efforts to repair deficiencies in the current approach. I will then offer an example of my own from educational research, as to how a causal analysis based on physical causal reasoning and system dynamics methodology might proceed.

Causality Reconsidered

In the social sciences causal inference is currently addressed through what is termed the counterfactual approach. The counterfactual approach recognizes an event to be the

cause of another if it can be shown that its presence is necessary for the second event (the effect) to have occurred. This requires demonstrating not only the condition: If X, then Y. It further requires demonstration of the counterfactual condition: If not X, then not Y.

The counterfactual approach is operationalized by experiment, or failing that, by analogous quasi-experimental designs. The designs vary in complexity, and particularly in the more recent work one often sees very sophisticated statistical models. In econometrics one speaks of simultaneous equation models. In sociology and political science one hears frequent references to path analysis and path models, and most recently to structural equation models. In sociology the methodology of structural equations was introduced in the 1960s by sociologists such as Blalock (1972) and Duncan (1975), and the new technology soon dominated the field, such that "few within sociology questioned the fundamentals; causal inferences followed automatically from structural equation models" (Berk, 1988, p. 155). The field of political science followed close behind (see for example Achen, 1986).

Criticism of this approach to causality has been growing for some years. D. A. Freedman is one of a number of statisticians who have pointed out that "you can't come close to checking the assumptions [of the models]" (1987b, p. 210), and that in any event the causal inference comes not from method but from theory. Recently, Lawrence Mohr (1996) has proposed an alternative approach, one that addresses certain shortcomings in the counterfactual approach, and in addition provides a credible rationalization for causal explanation in the absence of the conditions required for it. Mohr's proposals are subject to controversy. I accept them as Mohr has presented them, and proceed to show how they might be used. I urge the readers to investigate his claims for themselves. I have drawn heavily on Mohr's work, and I assume full responsibility for any errors and misinterpretations that may have occurred in the process. All references are to his 1996 work, unless otherwise noted.

Overcoming Weaknesses in the Counterfactual

The counterfactual definition states that "X was a cause of Y if and only if X was a necessary condition; if not X, then not Y" (Mohr, 1996, p. 27). Mohr draws on the philosophical literature to show that the counterfactual fails on either the "if" condition or the "only if" condition in various ways. In particular there are four widely recognized technical problems that pose severe challenges. To resolve these difficulties, Mohr introduces a modification that he calls the "factual approach," with an accompanying concept of the factual cause. Here is Mohr's definition of factual cause: "X was the factual cause of Y if and only if X and Y both occurred and X occupied a necessary slot in the physical causal scenario pertinent to Y" (ibid.). As Mohr points out, it is very close to the counterfactual definition.

Mohr then explains how and why this differs from the counterfactual approach:

> For the most part, the standard counterfactual definition captures the spirit of factual causation well. The shift from a necessary event, such as X, to a necessary slot, or function, is in most instances a minor one. The basis of the claim that something is necessary, however, and the basis of the recognition that it is indeed X that fills a certain slot are inaccessible to the standard counterfactual definition. The modified version, on the other hand, has ready, valid access to these claims and thereby is able to manage well the four technical challenges on which the standard definition was seen to founder. (ibid.)

Here is a brief example that Mohr gave in another work, of an application of the factual causal approach:

> Let us say . . . that the legislature raised the speed limit from 55 to 70 miles per hour and the highway death rate increased. Was the legislature's act a cause of the increase in the death rate? Surely it was not a physical cause. The new speed limit did not make people die, in the manner of billiard balls colliding on a table; in fact, it did not even make them drive faster. . . . In the true, physical causal scenario leading to the increase in the death rate in this instance, it may well be that the speed limit had to be raised in order for the additional deaths to occur. Because it was the legislature that raised the speed limit, the legislative act was a factual cause. (1995, p. 264)

The concept of physical cause Factual causality is one of two conceptions of causality that according to Mohr we use in social science. The other is physical cause. Physical cause is a relation between events in the natural world, as in "The hurricane tore the roof off of my house." Mohr goes on to describe it this way: "The first and the fundamental sense of causation is the physical. An instance of physical causation occurs when one object (large or small) contacts another in such a way that we can recognize a relation between a force and a motion (where by motion I mean motion other than what was to be expected from the law of inertia). The classic example is billiard balls on a table" (1995, p. 263).

Mohr acknowledges that this is in direct disagreement with David Hume's 18th century arguments, and mounts a lengthy argument in opposition to Hume (see 1996, chapter 2). The physical concept is important to Mohr's reasoning for two reasons. The first is that it is necessary to his idea of factual cause. The second is that it leads to a way to think about causality (which Mohr calls physical causal reasoning) that can be used to infer causality without resort to the counterfactual. More about that later.

Extending causal analysis to behavior Finally, Mohr has extended the concept of physical cause to cover intentional human behavior. Consider the problem. Even accepting that physical cause can be convincingly inferred, there has been no way of demonstrating a mechanism of physical cause between human intent and an event presumed to be the result of a behavior caused by that intent. Mohr claims that he has identified that mechanism. Again I leave it to the readers to decide the validity of this claim for themselves, and proceed here as if it has been established. What it does is extend the legitimate scope of physical causal reasoning to include the effects of human decisions and actions—subject matter that makes up the large part of the concern of the social scientist.

Increasing the Options for Social Science Research

The factual approach is based on determining whether a cause fills a necessary slot. The key term is necessary, meaning that if not X, then not Y. It is a variant of the counterfactual definition. It follows that reasoning about cause within the factual approach will be in terms of what would have happened if X had not occurred. Thus, "factual causal reasoning prominently includes experimental and quasi-experimental designs" (Mohr, 1995, p. 267). These designs may well be desirable where technical problems with the counterfactual are not anticipated, and where there is adequate data.

However, there are a great many questions in need of answers where those requirements cannot be satisfied. Let us revisit Mohr's definition of the factual cause. He states:

> [The definition of factual cause] may be elaborated more rigorously as follows: From our knowledge of physical causation, chains of causation, and the physical configuration of the world at the time of X, we know that at least k instances of physical causation resulting in various outcomes Y_i, i = r,... , k, had to occur just as they did in order for Y to occur just as it did. Either X was the physical cause of one of the Y_i or the implied alternative to X, call it "not X," would have been the physical cause of not Y_i. (To clarify the latter, if X is opening the shutter, for example, and Y_i is entering the air space between the shutter and the window, then opening the shutter [X] was not literally the physical cause of the ball's entering the air space [Y_i], but the closed shutter ["not X"] would have been the physical cause of the ball's being diverted from that air space [not Y_i].)" (p. 28)

Again we see that: (1) the factual definition rests on a physical causal scenario; (2) that scenario describes a process from a factual cause X (the shutter is open when a ball is thrown at a window) to a factual effect Y (the ball breaks the window). Ordinarily, one would collect many instances of the ball thrown with the shutter open, and a comparable number with the shutter closed. Taken together they would demonstrate that the ball breaks the window if and only if the shutter is open.

But the factual definition is also rooted in the chain of events itself. The X to Y scenario is available and accessible. Within the factual definition, we can examine the series of events directly to determine whether the shutter indeed occupied a necessary slot. The scenario would (and could) stand alone, a physical causal chain without a counterfactual. Even though the problem is stated as a factual cause, the data are available for analysis by other means—as a process, perhaps.

But what about that series of events? One may, in fact, have only intermittent quantitative data and some descriptive information, including the informal impressions of a number of witnesses. Let us say that what we have is sufficient to describe the events that link the factual cause to the factual effect. Is that description just storytelling?

Causal analysis with an N of 1, the modus operandi System dynamicists are fond of pointing out that there is plenty of information out there, if we can just find a way to use it. The great impediment to the existence of any alternative to the classical research design would seem to be that we apparently have no definition of causation that does not rely upon a counterfactual. Berk (1988), in acknowledging the weaknesses of quasi-experimental studies, observed that "it is the only apparatus we have, or at least the only one that is sufficiently developed" (p. 167). Berk is not entirely correct. There is at least one approach that is capable of demonstrating causality without recourse to a counterfactual. Scriven (1976) calls it the "modus operandi method." It might well prominently feature causal chains, but it is not attached to a quantitative design; it stands alone.

The modus operandi (MO) works as follows. An event, call it Y, has occurred, and we have compiled a list of several possible causes, X, U, V and W. Each of these suspected causes will leave a distinct trail, called its "signature," if it is present. This signature will consist of a mechanism or a known chain of events, or some other occurrences in addition to Y that are logically linked to the presumed cause. In Scriven's words, "The MO of a particular cause is an associated configuration of events, processes, or properties, usually in time sequence, which can often be described as the characteristic causal chain (or certain distinctive features of this chain) connecting the cause with the effect" (1976, p. 105).

The task is to sort through the evidence to see whether the signatures of one or more of the suspects are present, and which are not. Basically, the job is one of pattern recognition, and centers on discovering how many, if any, MOs of causes can be identified. Scriven offers the following procedures (1976, p. 106):

(i) Check for the presence of each A [i.e., suspected cause]. If only one, that is the cause.

(ii) If more than one is present, check for complete MO's. If none, then none of those A's was a cause.

(iii) If only one MO is complete, the A with which that MO is associated is the cause. If more than one complete MO is present, the associated factors are co-causes.

The MO is no arcane curiosity. It is currently in wide use, as Mohr tells us: "This [modus operandi] basis of determining causality is relied upon heavily in many areas, such as detective work, cause-of-death determination, medical diagnosis, and troubleshooting in connection with machinery, as in auto repairs" (Mohr, 1995, p. 262). Some social scientists also—anthropologists and historians come to mind—routinely make use of it. The modus operandi is more than just a convenient ad hoc device to be used in lieu of more formal methods, or at least so both Scriven and Mohr believe. It is itself a methodology. For Mohr, the basis of the modus operandi method is "physical causal reasoning" (1996, p. 116).

Physical causal reasoning Central to Mohr's argument for the inference of physical cause is his idea of physical causal reasoning. Here is my interpretation of that argument. Our fundamental understanding about cause is physical (based originally on sensations of forces upon our own bodies and generalized to the environment, see chapter 3), and because of this we are convinced by physical imputations of cause. We "know" that the reason a billiard ball rolls is because the cue ball smashed into it. It is largely due to long-standing philosophical arguments to the contrary that we are led away from that understanding. Consequently, if one were to come up with a sound defense of physical cause, that "natural" reasoning would be restored to credibility and acceptance, without discrediting counterfactual reasoning. A sound defense of physical cause is what Mohr has tried to achieve, drawing on the philosophical literature and physiological research to do so.

Because the concept of physical causal reasoning is at the core of Mohr's argument, I quote him extensively here. "Physical causal reasoning is entirely different [from factual or counterfactual reasoning]," writes Mohr. "There are no hypotheticals or counterfactual conditionals involved, such as 'if not X'; no comparison is necessary. Physical causation has to do with a force and a motion, and physical causal reasoning proceeds by showing that it was indeed force X that produced motion Y" (p. 102).

Mohr is aware that convincing his critics will not be easy: "Many have suggested to me at this point in the argument that physical causal reasoning is at bottom counterfactual reasoning—that we know the motion was the effect of the force because we know that if the force had not impinged on the object, the motion would not have occurred" (ibid.). Mohr answers that "This position is highly problematic," and responds with two major points (pp. 101-102). The first focuses on weakness in the counterfactual. "First of all,

the traditional counterfactual is not always true when X is a cause, and when the counterfactual is not true we can often detect the causation anyway. Counterfactual reasoning in those instances therefore cannot be the source of the correct causal inference."

At this point, Mohr interjects a reference to an example that appeared earlier in the book. The example is of an instance where Annie is about to stumble on a rock, and Manny and Fannie both yell for her to jump. Manny's yell is very loud, and so Fannie's soft-voiced warning is drowned out. Annie jumps. The counterfactual definition fails on the "only if" condition, because if Manny had not yelled, Annie would have heard Fannie and jumped anyway. Mohr goes on:

> Those who want to see a counterfactual behind every causal inference should recognize here that, in spite of the fact that there is no way we could have been properly informed by the pertinent counterfactual reasoning—and in fact would have been completely misled by it—we were unquestionably able to come to the correct causal conclusion anyway. Furthermore, there is a reason that we were able to do so, and this reason should be confronted; it should not be ignored for convenience.

Mohr's second point directly contrasts counterfactual and physical reasoning:

> Second, the counterfactual way often is simply not the way we reason, even when, as is usually the case, the counterfactual happens to be true. Suppose you stub your toe and cry out and sit down to hold it, and I say, "Why are you holding your foot?" and you say, "Because I stubbed my bare toe," and I then say, "How do you know it is because of that?" You would not say, "Because I know that if I hadn't stubbed my toe I would not be sitting down and holding it." It may be true that you would not be holding it—and then again there is some possibility that it actually is not true—but in any case that was not your reasoning process. You never stopped to think about that; it is not the way you arrived at your knowledge of why you are now nursing your toe. You arrived at your knowledge by recognizing certain feelings and desires: you felt a searing pain coming from the region of your toe as it encountered the iron leg of the bed and subsequently felt the desire to hold it because, for the moment, it *hurt!* To deny this is to risk being caught in a certain fanaticism or dogmatism, whereas to accept it is to begin to broaden the issue of causation in a potentially productive direction.

Is system dynamics reasoning so different? Mohr shows no hint of awareness of system dynamics. Nevertheless, the physical causal scenarios which underwrite his concept of factual cause lend themselves admirably to the system dynamics approach. Moreover, the concept of physical causal reasoning does not appear to be a strange concept to system dynamicists. Consider this resolute promotion of "operational thinking" that I ran across in my Stella documentation some time back (High Performance Systems, 1996, pp. 25 ff.).

> A prestigious economics journal contained a model that was designed to forecast milk production in the US . . . . The equation states that the dependent variable (Milk Production) is a function of a set of macroeconomic variables . . . . Clearly, the equation does not purport to represent how milk *actually* is produced. For nowhere in the expression do we see any *cows* . . . . Milk production is the product of Cows and milk per cow (per time). *This* is operational thinking . . . . By thinking in terms of how a process or system really works (i.e., its "physics"), we have a much better chance of understanding how to make it work better! This is what Operational Thinking does for you.

This reads to me very like something akin to physical causal reasoning (the physics of the process). And having made their case, the authors seem to imply that operational thinking is more basic than that underlying the (presumably counterfactual) logic of the cited equation, but that somehow people are being diverted from the realization:

"Operational Thinking appears to be very difficult for people to embrace. So deeply ingrained is correlational thinking that people (especially adults!) do not naturally think in an operational way" (p. 27).

Given the description of operational thinking, one is led toward the conclusion that the modus operandi approach has already been integrated informally into the system dynamics methodology. The basics of the modus operandi method are used by system dynamicists, as Forrester's (1991) comments indicate: "With regard to the use of data, system dynamics operates more like the engineering and medical professions, and less like practices in economics" (p. 25). Engineers and physicians are among those identified by Mohr and Scriven as practitioners of the modus operandi method.

Scriven (1976) argued for a much greater role for the MO in research: "I believe that the main thrust of efforts towards sophistication [in methodology] should now turn from the quasi-experimental toward the modus operandi approach" (p. 108). One of the problems, Scriven believed, was that "scholars whose field requires them to depend on the modus operandi approach often cannot articulate it well" (p. 102). Worse, those social scientists who have done this type of analysis have often done it "with a degree of informality that leaves it in the category of anecdotal evidence" (pp. 108-109). To my knowledge no one since Scriven has pursued the objective of improving the MO, but social scientists have been at work attempting to formalize various nonexperimental methods. Some of these efforts appear to be moving toward a systems view (see the discussion of logic models in Yin, 1998). It seems to me that system dynamics is poised to make a substantial contribution in this respect. If Scriven's suggestion strikes a chord, then perhaps his decades old plea has much to commend it to system dynamicists, and a sound philosophical foundation for physical cause would appear as important for system dynamics as it is for the modus operandi.

A Simple but Nontrivial Example

I will illustrate my interpretation of Mohr's approach to causal analysis with a straightforward problem drawn from educational research, and analyzed with a simple but nontrivial model. The problem concerns the causes of a decline in placement exam scores for local high school graduates applying to a community college. I propose to explain the decline by examining the physical causal scenario with a model of a school district, linking the results to reality by matching the model result to an empirical pattern.

Description

Beginning in the late 1970s the state of Florida introduced a minimum standards educational reform that endured throughout the 1980s. The reform was "tough" in the sense that it resulted in very high rates of retention in grade, and boasted a test to be passed in order to graduate. The reform was also marked by a high dropout rate. However, the reform did not produce the expected improvements in student performance. This was certainly the case in Dade County, Florida.[2] The Dade County Public School district (DCPS), the state's largest, informally withdrew its cooperation in 1987, and the reform was formally ended in 1990, accompanied by the general assessment that it had failed to improve public education in the state.

"Operational Thinking appears to be very difficult for people to embrace. So deeply ingrained is correlational thinking that people (especially adults!) do not naturally think in an operational way" (p. 27).

Given the description of operational thinking, one is led toward the conclusion that the modus operandi approach has already been integrated informally into the system dynamics methodology. The basics of the modus operandi method are used by system dynamicists, as Forrester's (1991) comments indicate: "With regard to the use of data, system dynamics operates more like the engineering and medical professions, and less like practices in economics" (p. 25). Engineers and physicians are among those identified by Mohr and Scriven as practitioners of the modus operandi method.

Scriven (1976) argued for a much greater role for the MO in research: "I believe that the main thrust of efforts towards sophistication [in methodology] should now turn from the quasi-experimental toward the modus operandi approach" (p. 108). One of the problems, Scriven believed, was that "scholars whose field requires them to depend on the modus operandi approach often cannot articulate it well" (p. 102). Worse, those social scientists who have done this type of analysis have often done it "with a degree of informality that leaves it in the category of anecdotal evidence" (pp. 108-109). To my knowledge no one since Scriven has pursued the objective of improving the MO, but social scientists have been at work attempting to formalize various nonexperimental methods. Some of these efforts appear to be moving toward a systems view (see the discussion of logic models in Yin, 1998). It seems to me that system dynamics is poised to make a substantial contribution in this respect. If Scriven's suggestion strikes a chord, then perhaps his decades old plea has much to commend it to system dynamicists, and a sound philosophical foundation for physical cause would appear as important for system dynamics as it is for the modus operandi.

A Simple but Nontrivial Example

I will illustrate my interpretation of Mohr's approach to causal analysis with a straightforward problem analyzed with a simple but nontrivial model. The problem concerns the causes of a decline in placement exam scores for local high school graduates applying to a community college. I propose to explain the decline by examining the physical causal scenario, using a model of a school district, and link the results to reality by matching the model result to an empirical pattern.

Description

Beginning in the late 1970s the state of Florida introduced a minimum standards educational reform that endured throughout the 1980s. The reform was "tough" in the sense that it resulted in very high rates of retention in grade, and boasted a test to be passed in order to graduate. The reform was also marked by a high dropout rate. However, the reform did not produce the expected improvements in student performance. This was certainly the case in Dade County, Florida.[2] The Dade County Public School district (DCPS), the state's largest, informally withdrew its cooperation in 1987, and the reform was formally ended in 1990, accompanied by the general assessment that it had failed to improve public education in the state.

During and following the years that the minimum standards reform was in effect at DCPS, the Miami-Dade Community College (MDCC) issued several reports of their admission and enrollment statistics. These reports presented data on the annual percentage of DCPS graduates seeking admission who scored above the college's cutoff on their entry-level placement examinations (Belcher & Downing, 1990; Rich, 1992; Belcher, 1993). The data from these reports, supplemented by data from state publications (Office of Postsecondary Education Coordination, 1994-1996), yield a 12 year pattern, graphed in Figure 1. The graph shows that the percent of DCPS applicants who were above the cutoff on all placement exams peaked at 45.1 percent in 1987, then dropped to a low of 21.0 percent in 1995 before appearing to recover slightly in the following year. The MDCC cutoff point is one indicator of a standard for a qualified high school graduate. Those who meet or exceed it are qualified; those who do not meet it are not.

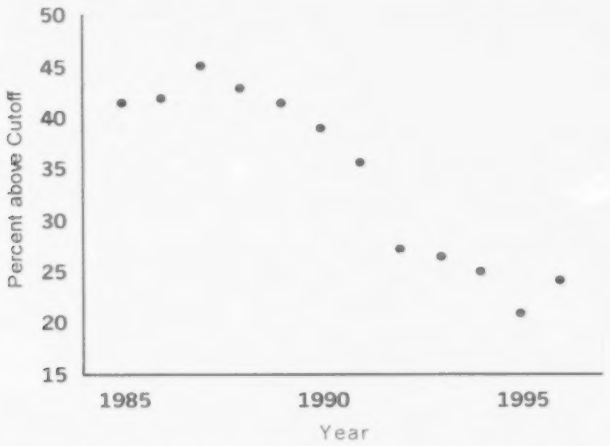

Figure 1. The percent of DCPS graduates applying for admission to MDCC who scored above the cutoff on all placement examinations.

The timing of the decline in qualified applicants coincides with the ending of the minimum standards reform in the school district in 1987. In view of the failure of the reform to improve standards, the most reasonable explanation for the decline is that the higher percentages of qualified DCPS graduates during the 1980s had been maintained by a high dropout rate (that is, poor performers dropped out in large numbers). The removal of the pressures to drop out caused a decline in the rate, increasing the numbers of unqualified students remaining to graduate.

It is important to make clear the relationship between the performance capability (percent of graduates qualified) of the DCPS graduate classes and the MDCC placement test results. A qualified graduate (QG) is defined as a student who, at graduation, is competent at the 12th grade level. The complaint that triggered the minimum competency reform was that many DCPS graduates were not performing at that level. The reform did not actually address that problem, but instead tried (via gate-keeper examinations and retention in grade) to guarantee a minimum 9th grade competency at graduation. This would presumably not affect the number of students who were performing at grade level

(i.e., they would have no fear of being retained or be intimidated by the testing). What is being hypothesized is that the number of below grade-level students who previous to the introduction of the reform policy were remaining to graduate, were under the reform dropping out in response to retention and other efforts to strengthen standards. Thus the DCPS percent of qualified students graduating (the percent QG) increased during the reform not because of any appreciable change in their numbers, but because as a group they became a proportionately greater part of the graduating class. Similarly, the percent QG declined after the termination of the reform because they made up a progressively smaller proportion of the graduating class.

The MDCC was committed to a policy of accepting all DCPS graduates who applied, and providing whatever remedial instruction was necessary. That was the purpose of the placement examinations; the cutoff score distinguished those applicants who were qualified from those who would need assistance. Since the qualified students (those performing at grade level) were not affected by the reform, their post-graduate choices were presumably unchanged, and the MDCC's share of them remained stable. It is the variation in the numbers of unqualified students graduating from DCPS that is of interest here. We know that the proportion of DCPS applicants meeting the MDCC cutoff varied considerably over the period 1985-1996. I make the reasonable assumption that the applications to MDCC from DCPS graduates unqualified at the 12th grade level varied in direct proportion to their numbers, and that this is reflected in the MDCC placement testing. If dropout was the cause of that variation, that should be reflected in the testing outcomes.

There are two paths by which dropout might have affected the percent QG. Two factors are known to have been present while the minimum standards reform was in effect—retention and dropout. Students were dropping out before the reform began and continued to do so. With the introduction of the reform, the retention rate increased sharply. Under those conditions an interaction between retention and dropout will cause the percent QG to vary. An increase in the retention rate results in an increase in the numbers dropping out due to the greater percentage of unqualified students enrolled (retained but not remediated). There are more students available to drop out even though the percent dropping out remains the same (is unaffected by the retention). This will be the case to the extent that prior conditions are the causes of both retention and dropout. The numbers dropping out will then decrease upon termination of the reform, with the cessation of the retentions, even while the dropout *rate* (computed as a percent of the At Risk student population) remains constant. As long as the dropout rate does not change, the variation in dropout numbers is a function simply of the changes in the enrollment. Retention does not directly cause dropout, though they clearly "vary together."

Or it may be that there is a change in the *rate* of dropout. This can occur if retention is a direct cause of dropout. If this is the case, then the dropout rate (computed as a percent of the At Risk student population) will increase in the presence of retention, due directly to the presence of the retention. That retention is a cause of dropout, if dropout can be shown to vary due to the retention rate, follows even given that other causes (e.g., low achievement) caused the retention that then caused the dropout.

Setting up a Strategy for Analysis

Before proceeding, it is desirable to eliminate other possible causes from consideration. The most obvious is that the reform had been successful, and the percentage of graduates scoring above cutoff dropped when it ended. The study of retention in DCPS elementary grades over this period undertaken by Morris and Hanson (1993), the opinion expressed in the report of the Governor's Commission on Educational Reform (1990), remarks by former DCPS superintendent Joseph Fernandez (Olson, 1990), and comments by Florida Commissioner of Education Betty Castor (Firing Line, 1992), all support the conclusion that the reform was a failure. A second possible cause might be that a change in the MDCC tests and/or procedures had occurred. There is no record of any changes prior to 1997, when MDCC joined a number of other Florida colleges in raising the cutoff scores. Third, changes in DCPS high school enrollments or graduates' preferences for colleges might affect the observed pattern, but no major shifts or notable changes were found to occur in the percentages applying to MDCC by high school, nor were there any shifts in enrollment or aggregate high school test scores. We are left, then, with the hypothesis that the decline occurred because fewer of the unqualified students dropped out after the reform was terminated.

Obviously, the reform was not the physical cause of the percent QG decline. Let us clearly state the hypothesis as a statement of factual cause: The fact that the reform was terminated caused the fact that the percentage of qualified graduates among the applicants to the MDCC declined. Were the data available, we might choose an appropriate quasi-experimental design, but they are not. We have only one district, in which all the schools were subjected to the reform, or were not, at the same time. There is not enough reliable time-series data on dropout available for a time series analysis. Moreover, the only available measure of student performance anywhere near the point of graduation is in the MDCC reports, and they cover only a 12-year period at the termination of the reform. In short, we have no counterfactual; we have what amounts to a case study. It will be necessary to resort to the modus operandi method or something like it, for the analysis.

There is another reason also, for preferring to examine the physical causal scenario. Studies similar to the one being developed here have sought unsuccessfully to unravel the relationship between retention and dropout through regular quasi-experimental analyses, and have failed due to the presence of one of the four aforementioned technical problems that plague the counterfactual, collateral effects (in the form of a spurious relationship). How the present approach handles this problem will be examined in the discussion.

Consequently, within the factual approach, we seek to learn whether the reform occupies a necessary slot in the physical scenario of events leading from the inception of the reform to the decline in QG. We have a hodgepodge of information of various quality, some of it quantitative and some not. We are able to identify a series of events which, when linked together, would result in a drop in graduate performance such as we see in the MDCC reports. That series of events is as follows. When the reform was introduced: (1) A basic skills standard ("the reform") was applied; (2) The reform caused a sharp increase in retentions; (3) The increase in retentions caused an increase in the number of AR students enrolled; (4) The increase in AR students caused an increase in the number of dropouts; (5) [Possibly] the retentions caused additional dropout; (6) The dropouts

were unqualified students who—in failing to graduate—led to an increase in the percentage of graduates who were qualified.

When the reform was terminated the process was reversed, and the percentage of qualified graduates decreased, in a similar chain of causal events. Making the reasonable assumption that the number of qualified graduates applying to the MDCC remained steady, this decrease should be reflected in the pattern of the college's entrance examination results.

That is our "theory" of the causal process. The MDCC data trace a complex 12-year pattern that was produced by a unique combination of events. If our theory can reproduce that pattern, the act of fitting such a complex sequence should constitute an adequate argument for its validity.

The Model

A model of the school district will be used to analyze the process we are trying to understand, and to see how that process can have produced the results we observe. A copy of the equation list is available on request.

Retention as a process The core of the model is the retention rate process. The retention rate pattern found to fit the aggregate empirical data on retention for a substantial number of American states is graphed in Figure 2. The process that produces this pattern is roughly as follows. A school's staff is given a standard against which to judge student performance, and instructed to hold back those students who do not meet that standard. Suppose that 50 percent do not meet the standard. Teachers assess students, and retain those whom they find have not met the standard. They will not find them all. They will identify, say, one-third of them. Then, teachers in the next grade will

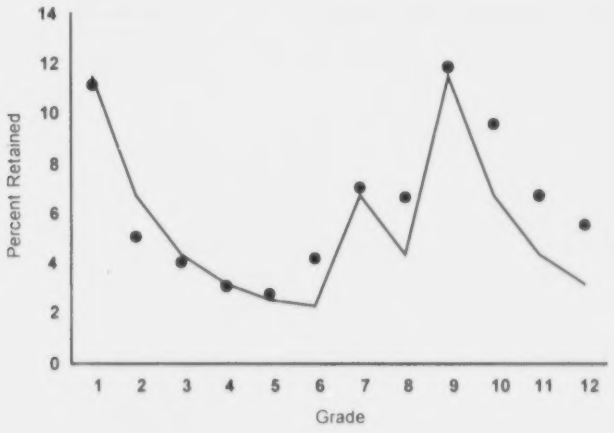

Figure 2. Retention Pattern across Grades. The solid line represents the exponential function. The symbols represent the average retention rate per grade of 11 American states in 1986. From Morris, 1993.

find one-third of the remaining, and retain them. This will continue until all those who are deficient in the standard are detected and retained, or until the situation changes, as when students go on to the next level. System dynamicists will recognize the process immediately. It is a simple goal-gap structure tracing out an exponential pattern of decay.

The same process is repeated at each educational level—elementary, middle, and senior. This is the pattern displayed in Figure 2, which shows the averaged data for 11 American states in the school year 1985-86. Morris (1993) found this pattern in state retention data for the school years 1979-80 and 1985-86. Updated data on state retention rates have been recently published covering 23 states, most across multiple years through the mid-1990s (Heubert & Hauser, 1999, pp. 139-147). These data show that the pattern identified by Morris persists over time and across more states. Karweit (1992) took notice of the peaks in the pattern: "Students are more likely to be retained at specific transition points, such as kindergarten or first grade (school entry) or Grade 6 (exit from elementary and entry into middle school), or Grade 9 (high school entrance)" (p. 1115). Gottfredson (1988) reported a similar observation.

Structure This retention process is embedded in a school system. The essence of a school system is a basic chain model, with an enrollment being channeled through 12 grades to conventional outcomes. The model processes students through the grades, and as it does so, they are divided into groups. The groups are created by dividing and then subdividing again a fixed enrollment entering the model at first grade (there are no other points of entry). The initial groups created upon entry are of at-risk (AR) and not-at-risk students. The not-at-risk group is essentially a placeholder group for determining the percentages. The at-risk group is further subdivided into retained and not-retained groups at a rate which defines the strength of the policy applied. Branching from the retained group is a further subdivision to accumulate the numbers remediated (if any). The remainder of the AR students are channeled on through as non-qualified graduates (i.e., students who graduate without meeting the standards) or dropouts. At the end of each of the educational levels one and two (elementary and middle) all retained but not remediated students are returned to the at-risk group and are again liable for retention at the next level.

The model produces three kinds of student outcomes. One outcome is the dropout. Dropout (which is ordinarily legally recognized at age 16) begins at the 9th grade for those students who are retained and continues through the 12th grade. The dropout for those at-risk students who are not retained is drawn from the 10th through the 12th grades. As a simplification, there is no dropout from the not-at-risk group, nor from the remediated group. The other two outcomes are qualified and unqualified graduates. Qualified is defined as meeting all current standards. It is assumed that all at-risk students remaining in school who are not specifically remediated will graduate unqualified. All not-at-risk students graduate as qualified. Students are remediated by being channeled into a remediation group from the retention group, and once remediated, remain so, graduating as qualified graduates.

Process The students proceed through the grades 1 – 12, one year at a time. Initially, in the pre-reform equilibrium, there is no retention, and no change in the size of the AR. A reform policy is then introduced, with the intent of retaining all who are performing below standard. The standard is introduced suddenly and applied equally to every grade.

The retention rate for those not meeting the standard is constant and one-third efficient. That is to say, one-third of those students who are at risk and not yet retained are retained, in each successive grade. The remediation rate is set to zero. This process generates the exponential decay pattern across the grades, conforming to the observed empirical pattern. Dropouts are drawn from both the At Risk and Retention chains beginning at grades 9 and 10—at equal rates if retention is not assumed to affect dropout, or with the students in the retained sequence subject to a higher rate if retention is assumed to affect dropout.

When a standards reform is introduced, it usually comes in on a wave of public support, and is initially imposed on all grades across the board. This means that in each of the 12 grades, one-third of the AR students are retained in the first year of the program, so that the number enrolled will increase sharply in a single year. Such sharp increases are reported by research. In her summary of that research, Karweit (1992) noted that retention in the Atlanta district quadrupled in 1981, the year following the introduction of a minimum standards reform there.

After that, one-third of the students at risk who were not detected in the first year will be retained the next year, and so on. All AR students are liable for retention at the beginning of each educational level, and by 12th grade many will have been retained three times. This too is supported by research. Shepard and Smith (1989) note that it is common for districts to limit retentions to one per level.

Results Using results from three runs of the model, I will focus here on the percent of graduates qualified (QG, the number qualified divided by all graduates, times 100). The two scenarios of interest, plus a "baseline" condition of no dropout, are displayed in Figure 3. The first run, the "No dropout" trace in the graph, represents the baseline run with the dropout rate set to zero. In the second run, represented by the "Normal dropout" trace, the dropout is set to represent the "normal" rate that prevailed prior to the introduction of the reform. The third run, shown in the "Ret-caused dropout" trace, doubles the normal dropout rate only for those students in the retention chain.

In the baseline trace, with no dropout in the scenario, we see the retention decay curve faithfully traced out by the QG sequence. This is the direct effect of the reform. It is the "pure signature" of the reform's effect on changes in the QG percentage. A little logic reveals that the initial peaks following the initiation of the reform are reflections of the abruptness of the introduction of the reform. Large numbers of unqualified seniors are retained all at once, and their graduation delayed a year. The effects from the initial retentions at seventh and first grades follow a few years later, after which the percent QG returns to the pre-reform equilibrium. What is of interest for this analysis is the repetition of the retention pattern in reverse, following the termination of the reform. This occurs because the reform is terminated abruptly. It shows that the retention rate alone can cause a temporary decline in percent QG. This effect was unanticipated, though obvious once the logic was apparent.

The percent QG varies directly with dropout. This shows up clearly when comparing the equilibrium levels before the introduction of the reform with and without a dropout rate (No dropout and Normal dropout). It follows that a change in the dropout rate will change the QG percent in direct proportion. Thus when normal dropout is added to the

model both the pre-reform and within-reform equilibria of QG increase. The distinctive characteristics of the retention signature remain, clear though somewhat diminished.

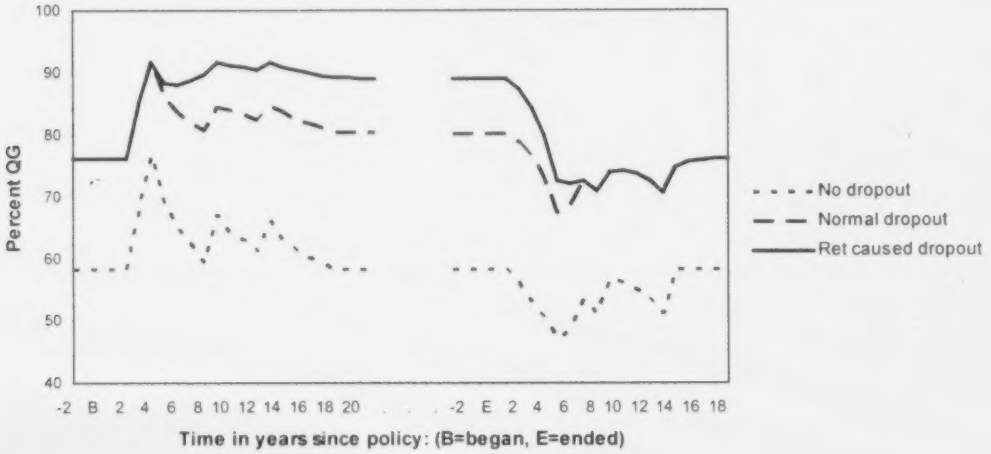

Figure 3. Model Output for the Percent Qualified Graduates. The "No dropout" trace represents the baseline condition where no dropout occurred. The "Normal dropout" trace represents the condition in which the dropout rate remained the same before, during , and after the reform policy. The "Ret-caused dropout" trace represents the condition where the dropout rate is doubled in the presence of retention.

When the assumption that retention causes dropout is introduced (Ret-caused dropout), doubling the dropout rate for retained students, during the reform, the percent QG is much increased and the retention disturbances much diminished. Again the retention characteristics remain, and in fact remain as they were under normal dropout conditions, except while the reform is in effect. The QG percentage at equilibrium is quite high.

Dropout, then, causes the magnitude of the change in the QG percent to vary, while the retention rate dictates the pattern of variation. The declines from the reform equilibrium mimicking the retention pattern are progressively greater as the dropout is increased, and the rebounds are smaller. This is particularly true of the first rebound following the major first decline, which all but disappears in the Ret-caused dropout trace.

Application to the Local Situation

To generate the results to be compared to the MDCC data, the model was adjusted to the settings and estimates specific to DCPS, based on data published by the district. The at risk student size was estimated from an equation using an elementary level Free and Reduced-price Lunch figure of 50 percent, about average for the period. Remediation was kept at zero. The existence of the triple-peaked retention rate pattern was verified with annual data published by the district, with peaks at 1st, 7th, and 9th grades, reflecting the district's dominant configuration at the time. Normal dropout (that is, the dropout rate prior to the introduction of the reform) was set to approximate 7.5 percent of the 9th through 12th enrollment, based on notes from conversations with the then Director of Testing for the district. Based on estimates and published data, the likely effect of retention on dropout was estimated to be 1.5 times the regular dropout rate.

The reform began in Dade in 1978, and ended in 1987 (the legislature formally ended the reform in 1990). That the reform was terminated quickly in Dade is verified by the rapid drop in retention rates observed in published district data (see Morris & Hanson, 1993). When the model is reset to reproduce the time span of the actual application of the reform at DCPS, the duration of the reform was brief enough that it ended before an equilibrium under the policy was reached. This is evident in Figure 4, where the lines show the model interpretations of the QG pattern across the full span of the years in which the reform was in effect.

There are no MDCC or other series data available to match to the time period between the start of the reform in 1978 and 1985, when the already displayed MDCC data begin. There is however, some corroboration from other sources. Informal interviews that I conducted in 1987-88 with district and school-site personnel concerning dropout prevention programs indicated that the retention rates increased a great deal in the first year of the reform. A local Grand Jury on dropout was convened in 1984, conforming nicely to the model results showing peaks in both the dropout (not shown) and QG rates at 1982.

The QG results from the two model runs were jointly fitted to the MDCC data pattern at the point of overlap (the three points at 1994-1996). These results are displayed in Figure 4, where the solid trace represents the run in which the dropout rate under retention increased, and the broken line the run in which the rate remained the same. From 1985, when the MDCC data begin, to 1993, after which the traces are the same, the average difference between them was just over 6 percentage points. The data (the symbols) fit the solid trace very closely. These results clearly indicate that the alternative in which

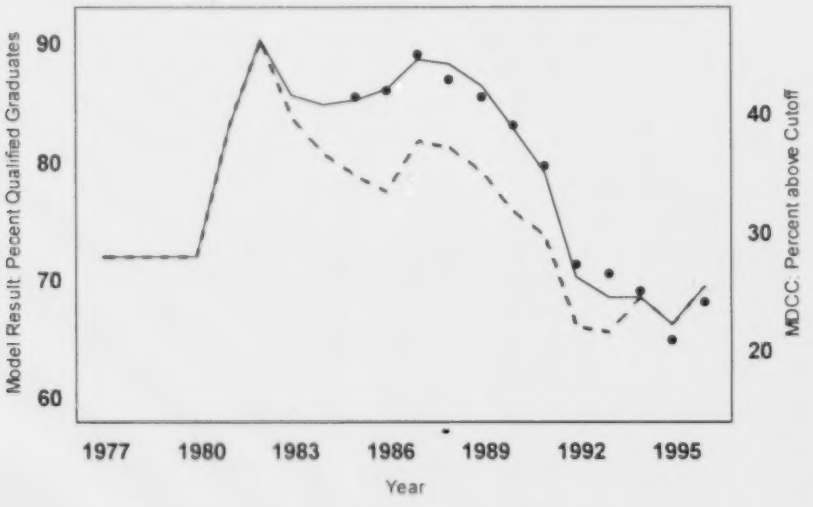

Figure 4. The MDCC Pattern Match. The broken line represents the model output for the condition under which the dropout rate is unchanged, and the solid line that in which the rate is increased by retention. The scale for the lines is on the left vertical axis. The symbols represent the percent of MDCC applicants who scored above the cutoff on the placement tests (right axis).

retention caused an increase in the dropout rate is the one that affords the best match. The model trace also conforms to the expectation that the percent QG range is considerably higher than that for the MDCC range. (Only a fraction of the district's graduates apply for admission to the college.) The fit of the model outcome to the empirical data is in fact an exceptionally good one.

In addition to the match of the MDCC data, another match was made to the pattern of a smaller fragment of data on the 9-12 grade dropout rate, a 5-year segment (1987-1991) from the school district, a source wholly independent of the MDCC data. Although the accuracy of dropout data is often questioned, this small segment coincides with the state's short-lived but well funded effort to stem dropout, when data accuracy was perhaps at its highest. The criteria for collection were changed in 1987-88 and again in 1994-95, making attempts to derive a longer series inadvisable. The datum for 1992-93 was omitted also, in response to warnings in district publications concerning data collection problems in the wake of 1992's Hurricane Andrew.

The dropout data is graphed in Figure 5. Again the solid trace represents the run with the increased dropout rate, and the broken trace the run in which the dropout rate remained the same for both retained and not retained. The round symbols represent the district dropout rate. The solid trace drops consistently from 9.5 percent of the 9-12 grade student body in 1987 to 7.4 percent in 1990, and then recovers. The rate for the solid trace is a little less than 2 percentage points above that of the broken trace at the maximum, indicating the magnitude of the direct effect of retention on the dropout. Since the data is from the system that is directly modeled (the school district), the scale for model output and empirical data are the same.

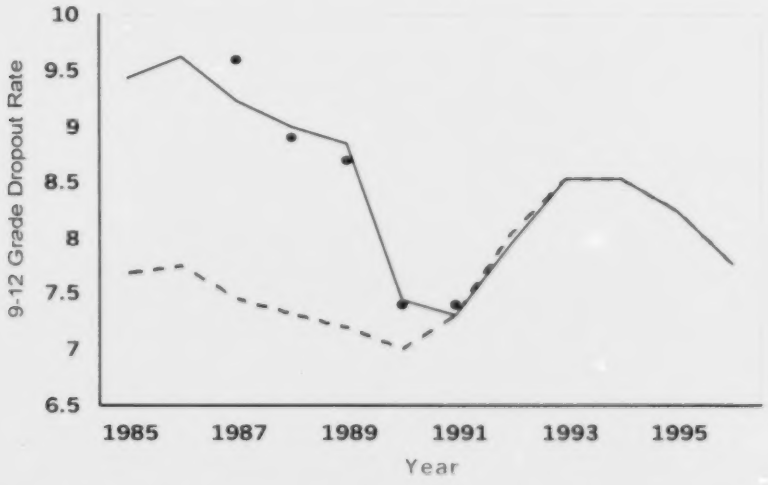

Figure 5: The DCPS Dropout Pattern Match The symbols represent the 9-12 grade dropout rate at DCPS for the years 1987-88 through 1991-92. The unbroken line is the model output when there is a retention-caused increase in the dropout rate, and the broken line represents the model output under the condition in which the dropout rate does not change. Data source: *DCPS District and School Profiles 1988-89, 1995-96*: Miami, FL.

The data in Figure 5 conform closely to the solid line, clearly indicating that the direct effect of the retention on dropout is real. This result is consistent with that displayed in Figure 4, and reinforces the conclusion that retention-caused dropout contributes to the variations in the MDCC entrance exam pattern.The fact that the model outcomes for both QG and dropout closely match the empirical patterns strengthens the argument we have advanced, and together the two fits bolster confidence in the model.

Discussion

Assessment of the Results

The physical causal approach We have framed the example problem as a factual causal statement, and from there sought to learn whether the reform filled a necessary slot in the decline of the percent QG in the MDCC data. With an N of one, we were required to examine the physical causal scenario from reform to MDCC, after the fashion of the modus operandi. With the model generating the physical causal scenario (filling the role of MO), we found that there were two ways in which dropout was caused to vary (and so reflected in the pattern match). In the first, the numbers of students dropping out increased/decreased as a result of the increase/decrease in at-risk students in the enrollment, caused by the increase/decrease in retentions. In the second, the dropout rate increased/decreased with the retention rate among those students retained at the senior level. Since retention is involved in both paths, the strong signature of the embedded retention pattern (from the goal-gap substructure) is clearly evident. The variation in dropout then affects the variation in graduates applying to the MDCC. Retention and dropout occupy necessary slots in the physical causal scenario and the reform (which was directly responsible for the retention) is a factual cause.

The next step was to evaluate the causal patterns generated by the model with respect to the empirical patterns. The pattern in the MDCC data is one that can only be produced by a unique combination of events. The MO method relies on identifying an expected signature to infer cause. Thus a close match of the data to an expected pattern implies that the associated cause is present. As Marquart (1990) has stated it, "The value of a pattern match is that the validity of the conclusions drawn from the data is strengthened if the pattern of results predicted by the theory is found in the data" (p. 94).

Pattern matches are assessed in various ways. Visual inspection is the most common, and we have already seen that the fit of the model outcomes to both data sets—the MDCC entrance scores and the DCPS dropout rate—is very close. Arguably the most salient aspect of the visual fits in this instance is the complexity. The more complex the pattern to be matched, the more convincing the match. The use of a system dynamics methodology makes the complexity issue especially relevant, because of the nonlinear nature of the models. In the present case, the model output's close fit to the details of the MDCC pattern is reassurance that our match reaches back through the physical cause scenario to the reform, our factual cause. The feedback in the retention rate generates a unique signature that dominates the output and is chiefly responsible for the fact that the model result fits the MDCC data almost point for point. This is persuasive support for the causal hypothesis.

Finally, applied researchers occasionally apply goodness-of-fit tests to pattern matches, analogous to the practice with statistical models. Aside from an alternate (non-visual) way of demonstrating that "the fit is good," it is not clear what such tests represent. In the interest of completeness, however, the correlation coefficients (Pearson's r) for the matches displayed in Figures 4 and 5 are as follows. The QG/MDCC coefficients (n = 12) are both very high: +0.993 for the run in which dropout rate changed in the presence of retention, and +0.959 for the run where the dropout rate was constant. This similarity in magnitude is to be expected. Both sources of dropout are causes, with the increase in the dropout rate adding modestly to the increase in dropout from the increased enrollment in that run. This result is in fact clearer in the graph. Of the coefficients for the dropout rate matches (n = 5), the one for the fit of retention-caused increase in the dropout is, at +0.977, very high, much higher than the +0.694 for the run where the rate remained the same throughout the reform. This, too, we know from the graph. The correlation coefficients tell us that the fits are very good, adding no new insights. The coefficients are not as dramatic as are the visual patterns, nor do they reveal anything of the conformity of the model outcomes to the complexity of the patterns; rather, they conceal details that are available visually.

Compared to a counterfactual approach That the retention-altered dropout rate contributes to the variation in the percent QG—as observed in the model outcomes—is clearly observed in both Figures 4 and 5. Our physical cause approach appears to have resolved a problem that the counterfactual approach cannot—that of spuriousness. Is retention a cause of dropout, or are both responding to some prior variable? This is a classic problem that has long plagued educational researchers. The problem has implications for policy, as Grissom and Shepard (1989) pointed out:

> Whenever high school dropouts and graduates are compared, it is always the case that a substantially larger proportion of the dropouts have repeated a grade. This observation has had little influence on school promotion policies, however, because it has been logical to say that grade retention is just another symptom of poor achievement, which is the real cause of dropping out. The purpose [of their 1989 study] was to analyze whether the retention decision itself increases the risk of dropping out. (p. 60)

In their study, Grissom and Shepard sought unsuccessfully to remove the ambiguity in the issue by way of a structural equation analysis, concluding that "causal modeling techniques can never produce unequivocal conclusions from correlational data" (ibid.).

This is one of the failings of the counterfactual approach that Mohr's factual approach repair-job claims to have rendered more tractable. He has observed that:

> If there is a concern about . . . spuriousness, [the] ordinary counterfactual approach becomes inadequate . . . . there is always the possibility of spuriousness based in a variable that was not measured, and perhaps not even imagined. A physical analysis can hold more hope of a definitive determination. It is only necessary to show by a physical analysis that X was or even can be a physical cause of Y. (pp. 100, 101-102)

In my example, the decrease in the dropout rate due to decreasing retention is a part of the physical causal scenario linking the termination of the reform with the change in MDCC placement exam outcomes. The model was set up to simulate the situation in which the dropout rate changes only among retainees in the senior high grades where the dropout occurs. This was then verified by showing that the sequence contributed to the accuracy of the matches to the empirical data patterns.

The physical approach also provides a good deal more information than does a counterfactual approach. The information about process is not required for a counterfactual approach, and might or might not be collected independently. Some of the results simply would not have been available at all. For example, two unexpected things were turned up by the model: (1) the effect of the exponential pattern of the retention rate; and (2) the effect of the abrupt start/stop of the policy. Both are important for understanding the timing and duration of the reform's effects. It is hard to see how either would ever be detected within an orthodox counterfactual design.

A system dynamics analysis, then, seems to fit well into Mohr's modified causal approach, and the approach appears to offer some advantages vis-à-vis the counterfactual. Questions remain, however, concerning how effectively such an approach can be defended as an explanation of the problem investigated (internal validity), and concerning the prospects for generalizing the results (external validity).

Internal Validity

Internal validity raises the question of whether the factors that we have identified are indeed the causes of the effects we have described. In other words, how convincing is our argument that the causes we have inferred are the right ones? The example that I have presented is an example of nonexperimental research. Case studies, and all other forms of nonexperimental research, have traditionally been supported by appeals to related research and general plausibility; the use of background knowledge adds a practical confidence to the study's findings. Research on events and processes that give added information on retention and dropout have here been cited and acknowledged. In like manner, local sources that have been consulted are referenced.

Since simulation has played a central role in the analysis, confidence in the model is also pertinent. Meadows (1980) has indicated that system dynamicists do not make an issue of internal validity in their models: "The system dynamics paradigm handles the problem of model validity qualitatively and informally. . . . [asking] Is the model sufficiently representative of the real system to answer the question it was designed to answer?" (pp. 36-37). Meadows lists three conditions which system dynamicists use to foster confidence in a model.[3] To the best of my knowledge, the model applied in the example given here meets them all.

Our traditional defenses, then, are in order, but is this the best that we can do? The approach employed here is patterned on the modus operandi as Mohr and Scriven have presented it. By consensus in the social sciences, internal validity is ordinarily established through quasi-experimental designs based on the counterfactual approach. The modus operandi does not depend on the counterfactual. Does that mean that the MO is weak as a basis for internal validity?

To address this question, consider first the technical aspect. Are the sophisticated path and structural equation model designs really any better at inferring cause than are other methods? If so, the advantage does not rest on their technical superiority. They have been subjected to severe criticism over the past two or three decades. Berkeley statistician D. A. Freedman is among the critics.[4] Freedman has taken the position that because the assumptions cannot be met or validated in social research, sophisticated statistical analyses such as path models cannot tell us much of anything about cause.

Freedman concludes his critique in this manner: "My opinion is that investigators need to think more about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models" (1987a, p. 125). Freedman does concede that regression methods may be of help in arguments about causation when used descriptively. He writes: "a regression equation—viewed as reporting a smoothed average of Y for each value of X—can be a link in a chain of reasoning about causes: the causal inference rides on the argument, not on the magic of least squares" (1987b, p. 209).

Freedman is saying quite plainly that social science approaches via any statistical modeling are just descriptive support for theoretical arguments of causal inference. This says nothing about the counterfactual approach as a philosophical persuasion, of course, but it would seem to place the defense of the quasi-experimental designs on an equal footing with that of the MO method. It comes down to a conviction about the underlying philosophical assumptions, "reasoning about causes."

Mohr bases his defense of the MO on physical causal reasoning:

> [The] criticism of the case study seems so devastating [because] there is no way of establishing causality except via comparison with an estimate of the counterfactual . . . . Modus operandi, however, appears to proceed by some different route. Only one instance is observed—one death in a car crash, for example. True, we may know from prior experience or research that heart attacks can cause both deaths and car crashes, and this knowledge may be important to us in some cases, but even then the prior experience or research may well not have used the regularity theory or factual causality to reach its conclusion and, in any case, neither is being used to determine whether a heart attack was indeed causal in this instance. The investigator does not explicitly consider crashes in which there was a heart attack and no heart attack, a death and no death. He or she simply looks for traces of a heart attack in the one case at hand. The whole idea of variables among which a correlation might be established seems irrelevant to this method. Is the modus operandi approach therefore weak as a basis of internal validity, or is there some elaboration of the idea of causation that shows it to be strong? I suggest that it is strong and that it is physical causal reasoning that makes it so. (p. 115)

This amounts to asserting that the internal validity rests on philosophical grounds. In this vein, Mohr observes that "we sometimes have substantial confidence that the case made for a physical cause is valid, just as we have a great deal of confidence at times, although certainly not all of the time, in the conclusions of factual causal reasoning" (p. 117).

Mohr's position is that in social science we should and do use two concepts of causality: the factual and the physical. The factual approach is essentially the equivalent of the counterfactual approach, and it is at the same time defined in terms of physical causal scenarios, which may be independently investigated by the MO method. It follows that we have a direct relationship between the MO and the counterfactual that is logical, clearly defined, and (assuming that one accepts physical causal reasoning), philosophically justified.

External Validity

If we accept physical causal reasoning, it is not internal validity but external validity that is the main source of the case study's disadvantages. External validity addresses the question of how general are the conclusions from this approach. Sweeping generalizations of the kind found in the physical sciences are not possible in the social sciences. Nevertheless, there are routes open to limited but valuable generalization.

Mohr approaches the external validity question from the standpoint of the sample size of the study, and concludes that a large N study based on probability sampling holds an edge over the case study (N = 1) relying on physical causality and the MO method. The advantage is that one is able to generalize to a limited population with a statistically defined precision. However, it is limited by very serious restrictions: the sample must be truly random; the generalization is restricted to the population sampled; and further to that (past) time period at which the observations were made. As Mohr points out, the advantage is a thin one. It is so thin that for Mohr, everything considered, the large-sample and case designs balance each other out.

The countering advantage of the case study lies in in-depth understanding. Mohr defends the value of in-depth understanding by arguing that such knowledge does increase our ability to make good judgments and respond appropriately in circumstances similar to those studied. Here is Mohr's considered conclusion concerning the relative status of the two approaches:

> The case study—research using a sample of only one but that one treated in substantial depth and detail—is commonly considered to be inferior to large-sample research in terms of both internal and external validity . . . . I want to suggest, however, that these conceptions of the limitations of the case study as a research design are superficial and overdrawn. It is extremely important in this connection to see that we have no designs at all in social science that will accomplish the dual aim of research [internal and external validity] . . . with a high degree of assurance and reliability. All designs have quite serious limitations with respect to either one goal or the other. I will advance the view that when the extent of these limitations is recognized and we speak relatively, the case study is potentially an excellent vehicle for advancing both of the general goals and should not *in principle* occupy an inferior position to large- or small-sample research of any description. (pp. 108-109)

In another publication, Mohr makes the case for in-depth understanding in a somewhat different way: He writes that physical causal reasoning has an advantage in that it "emphasize[s] understanding the method or mechanism by which the causation came about and . . . the more thoroughly we understand the causal mechanism by which a treatment [or policy, or event] has affected an outcome in one case, the better the position we are in to know when a similar outcome will result in another case" (1995, p. 271).

Although Mohr seems unaware of the existence of system dynamics, this reasoning leads directly to an external validity role for the model. A system dynamics model is, above anything else, a vehicle for understanding causal mechanisms. I suggest that the model—by formalizing "the conditions" (following the caveat "other things being equal")—allows a measure of generalization (external validity). The system dynamics approach adds a new *dimension* to generalization—classes of casual mechanism. If I understand the idea correctly, this is one of the implications of the concept of "generic structures."

There are two sources of generality in my example of educational reform—the chain structure of the school district and the goal-gap structure of the retention process—to which a system dynamics model lends great advantage. A school district has a lot of structure imposed upon it. With minor modifications, the chain model accurately describes virtually every public school district in the nation. That general structure is strongly reinforced by the fact that the retention process is a common and well-understood exponential decay process, and one that is found to occur in school districts all over the country. We can have considerable confidence that, given a school district

with moderate to high retention, abruptly altered and unaccompanied by remediation (not uncommon conditions), we will see the same pattern in graduate performance, because the model is what is generalizable.

Concluding Remarks: Physical Cause and System Dynamics

I have heard it said that system dynamics is a contradiction, that it is a qualitative methodology that is highly quantified. To what extent is the modus operandi an early, underdeveloped attempt at what a system dynamics model does so well? In his 1976 article, Scriven expressed his hopes for the possibility of quantifying the modus operandi approach. I think that a system dynamics model is very much in the spirit of what Scriven had in mind, although in a manner quite different from what he envisioned.

This is an area ripe for creative research. Does system dynamics need a stronger, more explicitly rationalized philosophical basis, and if so, is this the right way to go? Physical causal reasoning strengthens internal validity. Mohr has shown us how to pose the problems (ask the questions, set up the research strategy) from within a systematic and well-defended rationale. I think this is the greatest potential of Mohr's ideas for contribution to system dynamics. For their part, physical causal reasoning and the MO method are in need of more tangible claims to external validity, not to mention the analytical advantages inherent in the powerful concepts of feedback and systems thinking. These are the strengths of the system dynamics methodology. The two approaches have much in common, and complementary strengths that should serve to their mutual advantage.

I close by raising a question for the reader to ponder. With respect to understanding causation, there is a major difference between the present and earlier times. Today, computer technology makes available a much better insight into causation and causal mechanisms than did the simple observations of regularity and time-order that underlie the counterfactual approach. Forrester has made a distinction between the causes that we encounter in everyday life, and the more intricate causal structure of complex systems: "[I]n simple systems we learn that cause and effect are closely related in both time and space . . . . We repeatedly learn to expect a close association between action and the result. In more complex systems, however, the cause of a symptom may lie far back in time and in a remote part of the system" (1983, p. x). From this statement it seems clear that under conditions of close and immediate association, one is easily led to think in terms of action then result, and no action, no result. Such comparisons seem to lead naturally to methodologies based on linear algorithms. However, when time and space separate cause and effect, it seems impossible ever to make the connection without examining the process itself.

Granted, once the complex system has been modeled, such remote causes as have been discovered can then be subjected to study, and that study will no doubt consist of repeated trials under controlled counterfactual conditions, but that will not explain how we came to know of them as causes. The question then is this: To what extent is physical causal reasoning necessary to the discovery and inference of remote (and sometimes counterintuitive) causes?

References

Achen, C. H. (1986). *The statistical analysis of quasi-experiments*. Berkeley, CA: University of California Press.

Belcher, M. J. (1993, January). *Prepardness of high school graduates for college: A statewide look at basic skills tests results for 1990-91*. Information Capsule No. 92-01C, Office of Institutional Research. Miami, FL: Miami-Dade Community College.

Belcher, M. J. & Downing, S. (1990). *Who's prepared for college? Results of a five year study of recent high school graduates taking Miami-Dade's basic skills placement tests*. Research Report No. 90-04R. Miami, FL: Miami-Dade Community College, Office of Institutional Research. (ERIC Document Reproduction Service No. ED 328 317)

Berk, R. A. (1988). Causal inference for sociological data. In N. J. Smelser (Ed.). *Handbook of sociology* (pp. 155-172). Newbury Park, CA: Sage Publications.

Blalock, H. M., Jr. (1972). *Causal inferences in nonexperimental research*. NY: W. W. Norton & Company.

Duncan, O. D. (1975). *Introduction to structural equation models*. NY: Academic Press.

Firing Line. (1992, January 16). *A report card for schools*. Telecast transcript #927, distributed by Southern Educational Communications Association, Columbia, SC.

Freedman, D. A. (1987a). As others see us: A case study in path analysis. *Journal of Educational Statistics*. *12* (2), 101-128.

Freedman, D. A. (1987b). A rejoinder on models, metaphors, and fables. *Journal of Educational Statistics*. *12* (2), 206-223.

Forrester, J. .W. (1983). Foreword. In N. Roberts, D. Andersen, R. Deal, M. Garret, & W. Shaffer, *Introduction to computer simulation: A system dynamics modeling approach*. Reading, MA: Addison-Wesley.

Forrester, J. W. (1991). *System dynamics and the lessons of 35 years*. http://sysdyn.mit.edu/people/jay-forrester.html

Gottfredson, G. D. (1988). *You get what you measure - You get what you don't: Higher standards, higher test scores, more retention in grade*. Center for Research on Elementary and Middle Schools. (Report 29). Baltimore: The Johns Hopkins University.

Governor's Commission on the Reform of Education. (1990). *Reforming education in Florida*. Tallahassee, FL: Florida Department of Education.

Grissom, J. B. & Shepard, L. A. (1989). Repeating and dropping out of school. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention*, (pp. 34-63). Philadelphia: The Falmer Press.

Heubert, J. P. & Hauser, R. M., Eds. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

High Performance Systems, Inc. (1996). *An introduction to systems thinking*. Hanover, NH: Author.

Karweit, N. L. (1992). Retention policy. In M. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., pp. 1114-1117). New York: Macmillan.

Marquart, J. M. (1990). A pattern-matching approach to link program theory and evaluation data. *New Directions for Program Evaluation*, No. 47, 93-107.

Meadows, D. H. (1980). The unavoidable a priori. In J. Randers, Ed. *Elements of the system dynamics method* (pp. 23-57). Cambridge, MA: MIT Press. (Reprinted by the Productivity Press.)

Mohr, L. B. (1996). *The causes of human behavior: Implications for theory and method in the social sciences*. Ann Arbor: The University of Michigan Press.

Mohr, L. B. (1995). *Impact analysis for program evaluation*. Thousand Oaks, CA: Sage Publications.

Morris, D. R. (1993). Patterns of aggregate grade-retention rates. *American Educational Research Journal, 30*, 497-514.

Morris, D. R., & Hanson, M. K. (1993). Consequences of reform: Retention rate fluctuations in Dade County. In S. Bacharach & R. Ogawa (Eds.), *Advances in research and theory of school management and educational policy* (Vol. 2, pp. 121-157). Greenwich, CT: JAI Press.

Office of Postsecondary Education Coordination. (1994-1996). *Readiness for college*. Tallahassee, FL: author.

Olson, L. (1990, May 16). Education officials reconsider policies on grade retention. *Education Week on the Web*.

Rich, J. C. (1992, Feburary). *Performance of public high school graduates on Miami-Dade basic skills tests updated for Fall 1990 and Fall 1991*. Information Capsule No. 92-01C, Office of Institutional Research. Miami, FL: Miami-Dade Community College.

Richardson, G. P. (1991). *Feedback thought in social science and systems theory*. Philadelphia: University of Pennsylvania Press.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual* (vol. 1, pp. 101-118). Beverly Hills, CA: Sage Publications.

Shepard, L. A., & Smith, M. L. (Eds.). (1989). *Flunking grades: Research and policies on retention*. New York: Falmer Press.

Yin, R. K. (1998). The abridged version of case study research. In L. Bickman & D. J. Rogg (Eds.), *Handbook of applied social research methods* (pp. 229-259). Thousand Oaks, CA: Sage Publications.

Notes

1. The dispute over the nature of causation dates far back in history, and that history has not gone completely unnoticed in the field of system dynamics. At least one system dynamicist, George Richardson (1991), has acknowledged that "there are a host of questions about causality in social science, including whether the concept has any scientific meaning at all" (pp. 7-8). However, he found it necessary for his purpose to assume those questions away: "I choose simply to presume that the concept of cause in the social and policy sciences has meaning, from which we can derive a meaningful idea of closed loops of circular causality" (ibid.).

2. The county's name (and so that of the school district) was recently changed to Miami-Dade. To remain consistent with the sources, I have kept the old name (Dade) throughout.

3. The conditions (Meadows 1980, p. 37) are as follows.

> 1.Every element and relationship in the model has identifiable real-world meaning and is consistent with whatever measurements or observations are available.
>
> 2. When the model is used to simulate historical periods, every variable exhibits the qualitative, and roughly quantitative, behavior that was observed in the real system.
>
> 3. When the model is simulated under extreme conditions, the model system's operation is reasonable (physical quantities do not become negative or exceed feasible bounds; impossible behavior modes do not appear).

4. Freedman's arguments were made the focus of a special issue of the *Journal of Educational Statistics* in 1987. Freedman's critique and the responses from researchers from a variety of social science disciplines are well worth the reader's attention.